An Interpretable Approach to Fake News Detection

Caio Brighenti
Department of Computer Science, Colgate University

April 17, 2020

Abstract

The text of your abstract. 200 or fewer words.

Keywords:

1 Introduction

Propaganda has long been a tool of political influence, but in recent years it has taken a new online form: fake news. Fake news, once a buzzword on the internet, is now at the center of global politics, one of the most common bigram in the lexicon of United States president Donald Trump. After the term gained prominence in Trump's presidential campaign in 2016, it exploded into public conciousness, earning the distinction of Webster-Collins' "Word of the Year" in 2017.[1] As the current presidential race unfolds, fake news has returned to the center of the conversation, with major social media companies facing scrutiny of their misinformation policy. This phenomenon is also not a distinctly American problem–investigate reporting both during and after the 2018 Brazilian president election demonstrated that more than XX% of news articles shared on the popular messaging service WhatsApp were fake news.[2]

Fake news is not only politically significant but also dangerously tempting. Studies have shown false content propagates faster through social media than real content.[3] Blatantly false or exaggerated rhetoric can even lead to violent action, as demonstrated by the "Pizzagate" incident in which a man stormed a D.C. pizzeria with an AR-15, having been convinced by false and unverified information that a pedophile ring operated out of the restaurant's basement.[4] Fake news can also be incredibly easy to create. In 2019, a group of researchers at the Allen Institute for Artificial Intelligence published text generation model able to produce fake news.[5] In a troubling conclusion, the researchers found that state-of-the-art fake news detection systems struggled more with identifying fake news produced by their systems than actual fake news.[6] Fake news is thus easy to create, spreads quickly, and is hard to detect, a dangerous combination making it a serious threat to civic society.

Given the danger that fake news poses, machine learning and natural language processing researchers have devoted significant attention to the problem of fake news classi-

1

2

3

4

5

6

fication. However, previous attempts at fake news classification overwhelmingly rely on highly complex models suffering from the "black box" problem. As a result, these models lack interpretability and do not allow us to reach new conclusions about the nature of fake news. In order to begin closing this gap, this research adopts an interpretable approach, with the overall objective of a producing a fake news classification model with comparable accuracy to state of the art models without compromising interpretability.

PARAGRAPH SUMMARIZING FINDINGS

2 Prior Work

Since the 2016 U.S. presidential election, fake news has been a frequent topic of natural language processing research. There are countless examples of papers approaching fake news classification or closely related problems, but from slightly different angles. This section summarizes the prior work in fake news detection, clarifying the different categories of methods. In general, any fake news detection model falls into one of three levels of granularity: 1) claim level, 2) source level, and 3) article level. These levels of analysis describe the response variable being predicted.

Claim level approaches attempt to determine whether a given claim, usually one or several sentences, is true or intentionally misleading. [7] Given that claim-level approaches must make a judgement based on only a short amount of text, researchers often adopt a fact-checking strategy This strategy, also known as "truth discovery," assumes that a sentence's claims can be gramatically isolated and checked against a database of established claims.[8] A natural application for claim-level models is social media, most commonly Twitter, where little is known about the author and only a very limited amount of text is available. However, claim-level approaches have serious limitations, often struggling with the complex sentences journalists or other writers typically employ.[9] Aditionally, they rely entirely on a complete knowledge base, which must be constantly expanded and updated, clearly a difficult task.

[7] Examples of claim-level approaches include ...

[8] strube p. 2

[9] strube 2

Source level approaches attempt to classify whether a speaker or entire news source consitently publish misinformation. The intuition behind these approaches is that speakers or sources that have published misinformation in the past are likely to continue to do so. An example of a source-level approach is the popular browser extension "BS Detector,"[10] which classifies articles on a fine-grained scale of veracity by checking the source's status in a database of news sources and their reliability. A source's history of misinformation can also be used as a predictor in claim-level or article level approaches. Kirilin and Strube, for instance, create *Speaker2Credit*, a metric of speaker credibility, and show how it can improve the performance of fake news detection models when used as an input.[11]

Article level approaches have received the most attention in the work on fake news detection. This is logical, given that fake news tends to take the form of articles, peddling misinformation in the article text while posing as a legitimate source. Article-level approaches also benefit from a rich list of predictors to choose from, including the article's content, author, source, metadata, claims made, as well as the social context around it, meaning information on how it has been discussed and shared on social media.

Accross all three levels of analysis, researchers overwhelmingly choose to use complex deep neural networks. Ajao et. al, for instance, use a "hybrid of convolutional neural networks and long-short term recurrent neural network models" to classify Tweets as true or false based on their text content.[12]

While deep learning approaches can produce highly accurate models that consistently succeed in identifying misinformation, they also suffer from a lack of interpretability. This is often referred to as the "black-box" problem, meaning that the inputs and outputs of these models are perfectly clear, but the steps that the model takes to reach the output are completely invisible. This has been identified as a limitation of the the work on fake news detection thus far.[13]

[10]

[11] Kirilin and Strube

[12] ajao et al

[13] shu, O'Brien

3 Methodology

4 Results

5 Discussion

6 Future Work